

Citation for published version:

Patel, M, Coles, S, Giaretta, D, Rankin, S & McIlwrath, B 2009, 'The role of OAIS representation information in the digital curation of crystallography data', Paper presented at IEEE eScience 2009, Oxford, UK United Kingdom, 9/12/09 - 11/12/09. <https://doi.org/10.1109/e-Science.2009.27>

DOI:

[10.1109/e-Science.2009.27](https://doi.org/10.1109/e-Science.2009.27)

Publication date:

2009

Document Version

Publisher's PDF, also known as Version of record

[Link to publication](#)

Publisher Rights

CC BY-SA

University of Bath

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

The Role of Representation Information in the Digital Curation of Crystallography Data

**Manjula Patel¹, Simon Coles², David Giaretta³,
Stephen Rankin³, Brian McIlwrath³**

¹UKOLN, University of Bath, UK

²EPSRC NCS, University of Southampton, UK

³Science & Technology Facilities Council, UK

5th IEEE International Conference on e-Science
9-11th December 2009
Oxford, UK



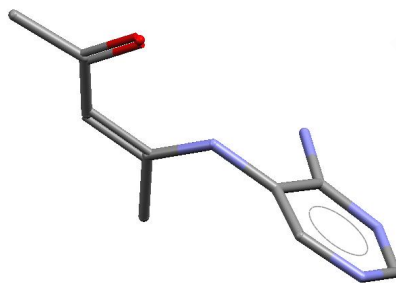
This work is licensed under a Creative Commons Licence: Attribution-ShareAlike 3.0
<http://creativecommons.org/licenses/by-sa/3.0/>

Data Deluge

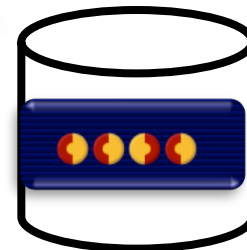
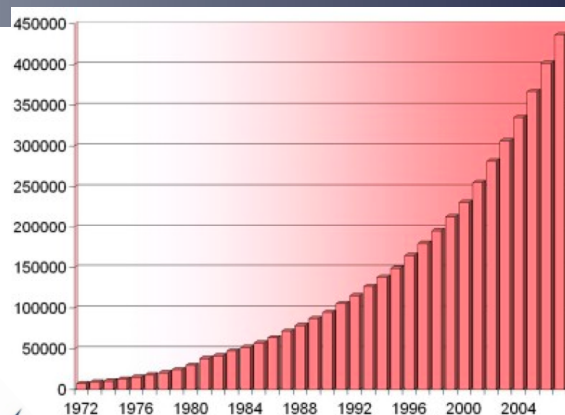
40 years ago a PhD student would determine about 3 crystal structures for their thesis – this can now be easily achieved in a day



35 million



2.5 million



0.5 million



'Few thousand'

The primary cause is the current data publication process, which is tied to journal articles and peer review

The Solution

research papers

Acta Crystallographica Section B
Structural Science
DOI: 10.1107/S0108768207010000

Serap Belli,¹ Simon J. Coles,² David B. Davies,³ Michael R. Haddon,⁴ Adam Kite,⁵ Thomas A. Mayes,⁶ Robert A. Shone⁷ and Aydin Ushak⁸

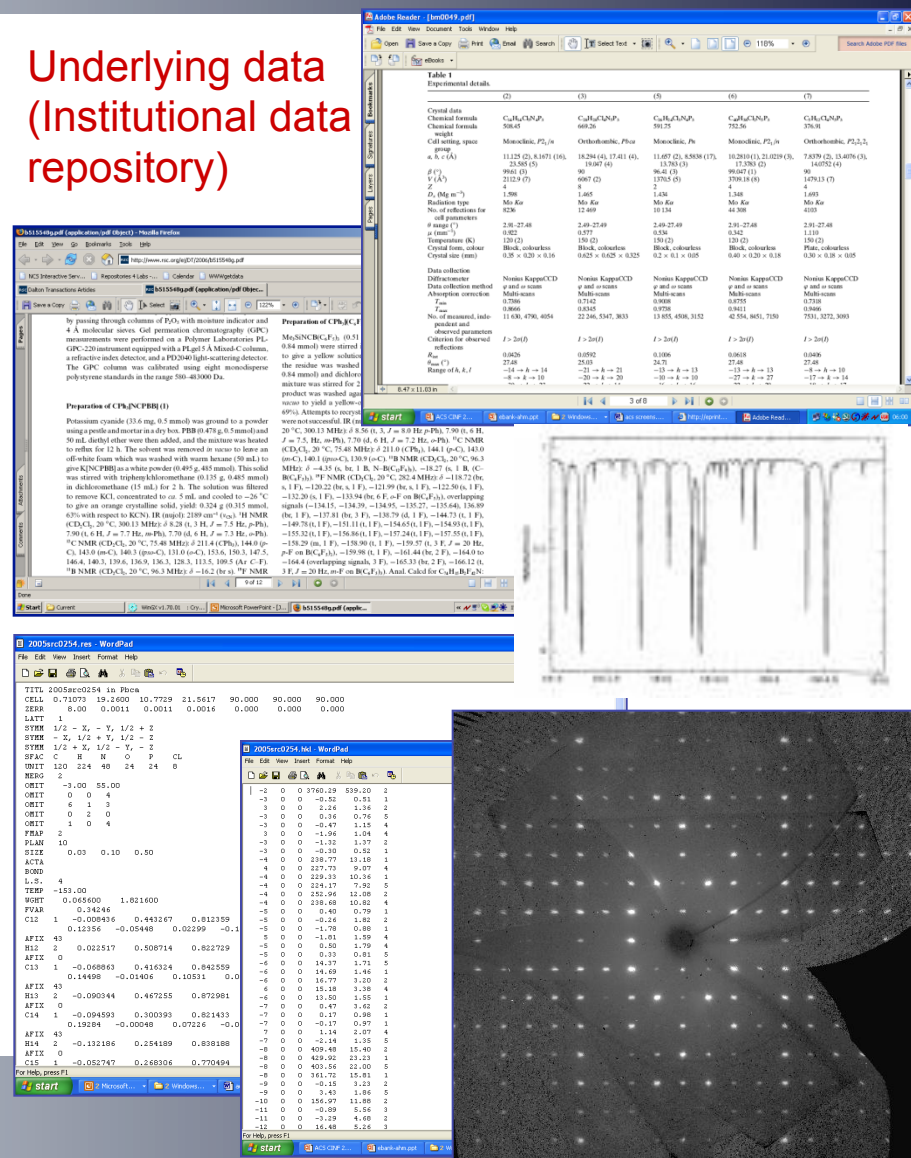
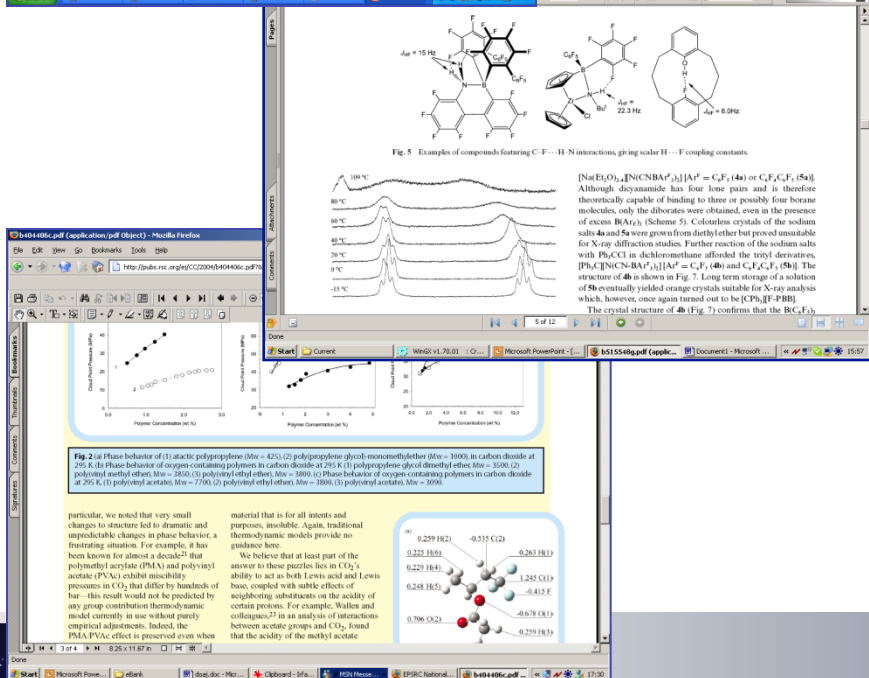
¹Department of Chemistry, Gates Institute of Technology, Gates Valley, Department of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, England, and ²School of Biological and Chemical Sciences, Birkbeck College University of London, Gower Street, 29 Gordon Square, London WC1H 0AP, England

Correspondence email: s.j.coles@ucl.ac.uk

A systematic study is presented on the products of aminolysis of $N_3P_3Cl_3$ (1) and $N_3P_3Cl_4$ (4) with dibenzylamine. Two series of mono- and disubstituted derivatives of compounds (1) and (4), namely $N_3P_3Cl_2[N(CH_2Ph)_2]$ (2) and $N_3P_3Cl_2[N(CH_2Ph)_3]$ (5) and $N_3P_3Cl_3[N(CH_2Ph)_2]$ (6) and $N_3P_3Cl_3[N(CH_2Ph)_3]$ (7) (where (2), (3), (5) and (6) are new structures), are investigated in order to determine whether steric or electronic effects prevail in the formation of dibenzylamine-substituted cyclophosphazenes. The influence of an electron-releasing group (i.e. phenyl) on the stereochemistry and degree of substitution of the product is analysed by comparison of the above two series. The difference in experimentally substituted endocyclic P–N bond lengths, Δ , is used as a measure of the degree of the electronic contribution, in combination with basicity constants, to quantify the degree of the electron-releasing capacity of the R group. In order to compare geminal versus non-geminal substitution, a difunctional secondary amine was used to form the compound $N_3P_3Cl_2[N(CH_2CH_2CH_2NMe)_2]$ (7) (reintroducing

Intellect & Interpretation (Journal article, report, etc)

Underlying data (Institutional data repository)



The eCrystals Data Repository

eCrystals UNIVERSITY OF Southampton

Home | About | Browse by Year | Browse by People

Logged in as Dr Richard A. Stephenson | [Manage deposits](#) | [Profile](#) | [Saved searches](#) | [Review](#) | [Admin](#) | [Logout](#)

2,2-trimethylenedioxy-4,4,6,6-tetrachlorocyclotriphosphazene

Sample Originator: A. Kilic^a, M. Odlyha^a, A. Uslu^a, David B. Davies^b and R.A. Shaw^b.

Data Collection: Mark E. Light^a, Simon J. Coles^a and Susanne. L. Huth^a

Structure Determination: Simon J. Coles^a, Michael B. Hursthouse and J.S. Rutherford.

^aGebze Institute of Technology
^bBirkbeck College
University of Southampton

C₃H₆Cl₄N₃O₂P₃

InChI=1/C3H12Cl4N3O2P3/c4-13(5)8-14(6,7)10-15(9-13)11-2-1-3-12-15/n8-10,13-15H,1-3H2

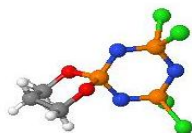
Identification Number: 10.3737/ecrystals.chem.soton.ac.uk/300

Controlled Keywords: cyclophosphazene, phase transition, variable temperature

Date Created: 28 March 2007

Deposited On: 21 Apr 2008 15:56

Deposited By: Dr Simon J Coles



Jmol

Available Files

Final Result

2005sjc0007.cif	11k
2005sjc0007.cml	4k
2005sjc0007.fcf	138k

Validation

2005sjc0007_checkcif.htm	9k
--	----

Refinement

2005sjc0007.res	5k
2005sjc0007_xl.lst	29k

Solution

2005sjc0007.prp	5k
2005sjc0007_xs.lst	44k

Processing

2005sjc0007.hkl	532k
2005sjc0007.htm	11k
2005sjc0007_0kl.jpg	91k
2005sjc0007_h0l.jpg	87k
2005sjc0007_hk0.jpg	79k

Data Collection

2005sjc0007_crystal.jpg	17k
---	-----

Other Files

2005sjc0007.doc	186k
2005sjc0007.inchi	1k
2005sjc0007.ins	4k
2005sjc0007.mol	2k
2005sjc0007.p4p	1k
2005sjc0007_ellipsoid.gif	

- Quick & simple to deposit
- Software tools
- Laboratory archive
- Community involvement
- 'Embargo' facility
- Structured foundations
- Discoverable & harvestable
- Number of file formats

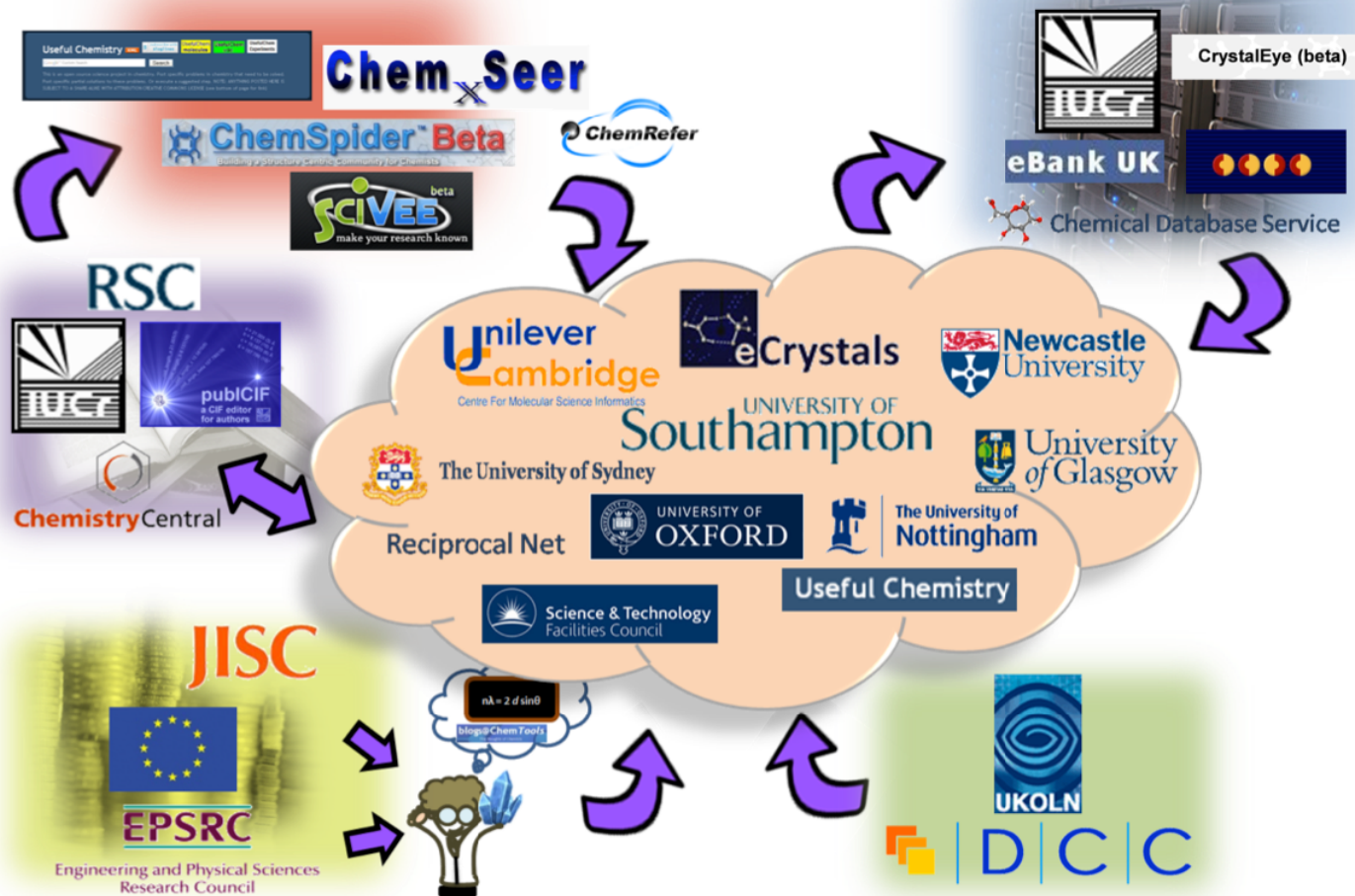


<http://ecrystals.chem.soton.ac.uk>


A Thorough Approach to Dissemination

- Using simple Dublin Core protocol (OAI-PMH)
 - Crystal structure
 - Title (Systematic IUPAC Name)
 - Authors
 - Affiliation
 - Creation Date
- Additional **chemical** information through Qualified Dublin Core
 - Empirical formula
 - International Chemical Identifier (InChI)
 - Compound Class & Keywords
- Specifies which 'datasets' are present in an entry
- Application Profile <http://www.ukoln.ac.uk/projects/ebank-uk/schemas/>
- DOI links <http://dx.doi.org/10.1594/ecrystals.chem.soton.ac.uk/145>
- Rights & Citation <http://ecrystals.chem.soton.ac.uk/rights.html>


Building a Federation of Repositories



Crystallography Data Commons



My Workspace | [cdlink](#)



[Home](#)
[Schedule](#)
[Announcements](#)
[Resources](#)
[Discussion](#)
[Chat Room](#)
[Wiki](#)
[Email Archive](#)

cdlink

[Options](#)

The cdlink Sakai site is a collaboration site for the development of repository requirements and frameworks for raw and derived crystallographic data.

Recent Announcements


[Options](#)

There are currently no announcements.

Recent Discussion Items

[Options](#)


First cut
(Simon Coles - Nov 12, 2008 7:11 AM)



Science & Technology
Facilities Council

[Home](#) | [Developer](#) | [Community](#) | [Download](#) | [Blog](#) | [Contact Us](#)

UNITE YOUR RESEARCH WITH

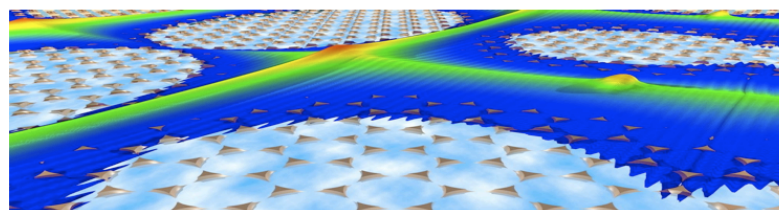


WHAT WILL IT ENABLE YOU TO DO?

- Access data anywhere via the web
- Annotate your data
- Search for data in a meaningful way e.g. taxonomy, Sample, temperature, pressure etc.
- Share data with colleagues
- Access data via your own programs (C++, Fortran, Java etc.) via the ICAT API
- Identify potential collaborations
- Utilise integrated e-Science High-Performance Computing and Visualisation resources
- Link to data from your publications

WHAT IS ICAT?

ICAT is a database (with supporting software) that provides an interface to all ISIS experimental data and will provide a mechanism to link all aspects of ISIS research from proposal through to publication.



WHY IS IT IMPORTANT?

The storage, retrieval and management of data is a major concern for all large scale facilities. For example, ISIS currently produces ~1TB of neutron and muon data each year and with the introduction of Target Station II, this rate of data collection is set to rise still further. The full value of these data resources will only be realised if they are easily searchable, accessible and reusable

ICAT DEVELOPERS WORKSHOP

All the presentations from the ICAT Developers Workshop at The Cosensers House, Abingdon along with notes from the meeting are available on the [meeting page](#)

- Generic publication & dissemination
- Domain & context
- Management
- Preservation Data for Crystallography Data
- Core Scientific Data Model

OAIS Background

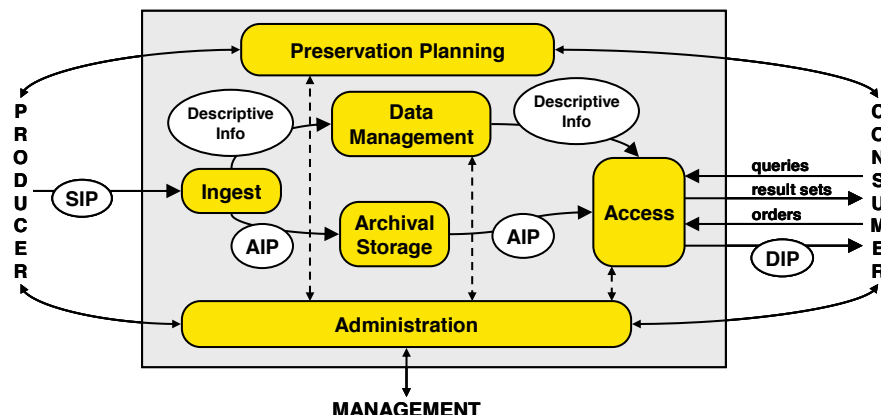
- OAIS -*Reference Model for an Open Archival Information System* <http://www.ccsds.org/documents/650x0b1.pdf>
- Development led by the Consultative Committee for Space Data Systems (CCSDS)
- Adopted as ISO 14721:2003
- “Open” refers to development of the model in an open forum
- Reference Model, not a blueprint for implementation
- Establishes a common framework of terms and concepts
- Identifies the basic functions of an OAIS
- Defines an information model
- Three major areas of influence:
 - Preservation metadata schemas
 - Architecture and system design
 - Conformance criteria for archival repositories

OAIS Definition and Selected Concepts

- **OAIS:** “An archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community”
- **Designated Community:** Community of stakeholders and users that the OAIS serves
- **Knowledge Base:** A set of information, incorporated by a user or system, that allows that user or system to understand the received information
- **Information Object:** Data Object + Representation Information
- **Representation Information:** any information required to render, interpret, use and understand digital data
- **Information Package:** Content Information + Preservation Description Information + Packaging Information (Submission, Archival and Dissemination Information Packages)
- **Preservation Description Information:** Provenance, Context, Reference, Fixity information

OAIS Functional Entities

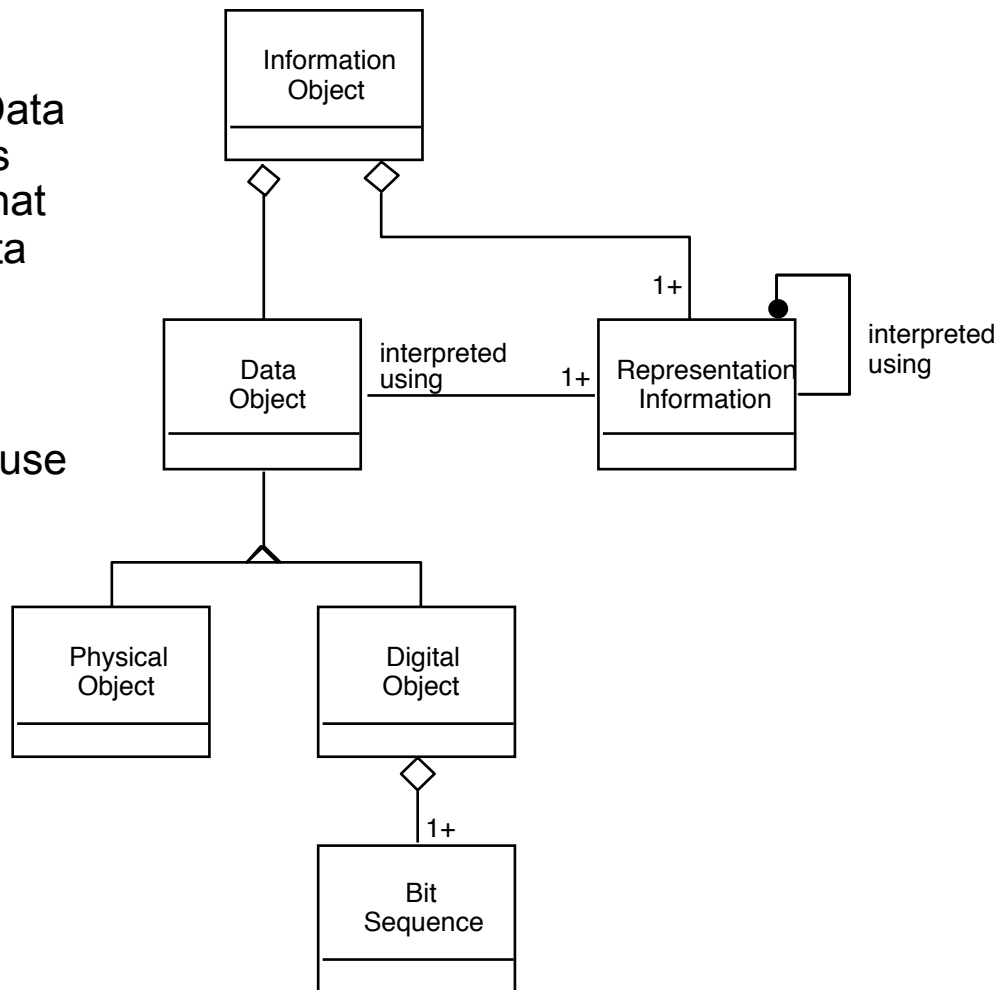
- **Ingest:** services and functions that accept SIPs from Producers; prepares AIPs for storage, and ensures that AIPs and their supporting Descriptive Information become established within the OAIS
- **Archival Storage:** services and functions used for the storage and retrieval of AIPs
- **Data Management:** services and functions for populating, maintaining, and accessing a wide variety of information
- **Administration:** services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis



- **Preservation Planning:** services and functions for monitoring the OAIS environment and ensuring that content remains accessible to the Designated Community
- **Access:** services and functions which make the archival information holdings and related services visible to Consumers

OAIS Information Model

- **Information Object** is composed of a Data Object that is either physical or digital, as well as the Representation Information that allows for the full interpretation of the data into meaningful information
- **Representation Information** is *any* information required to render, interpret, use and understand data



OAIS Representation Information (RI)

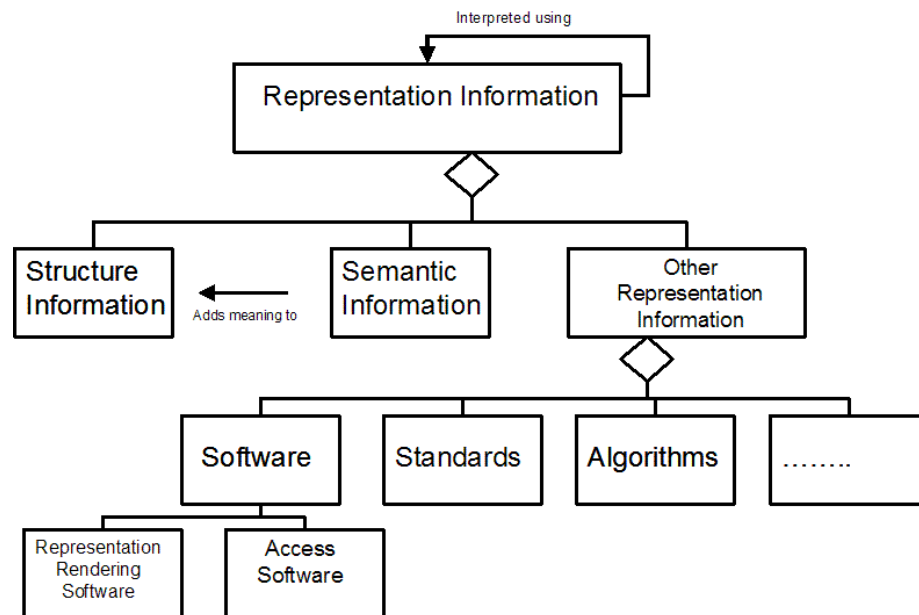
- Types of RI

Structure e.g. file formats for text, images, audio, moving images, datasets, 3D models

Semantic e.g. data dictionaries and knowledge organisation systems such as schemata, ontology, metadata vocabularies and thesauri

Other e.g. software, algorithms, standards, time dependent information, actions, processes

- RI is recursive in nature; using one element of RI in a meaningful manner may well require further RI, resulting in a RI Network
 - Recursion is terminated based on the designated community's knowledge base
 - Essential that RI itself is curated and preserved to maintain access to data



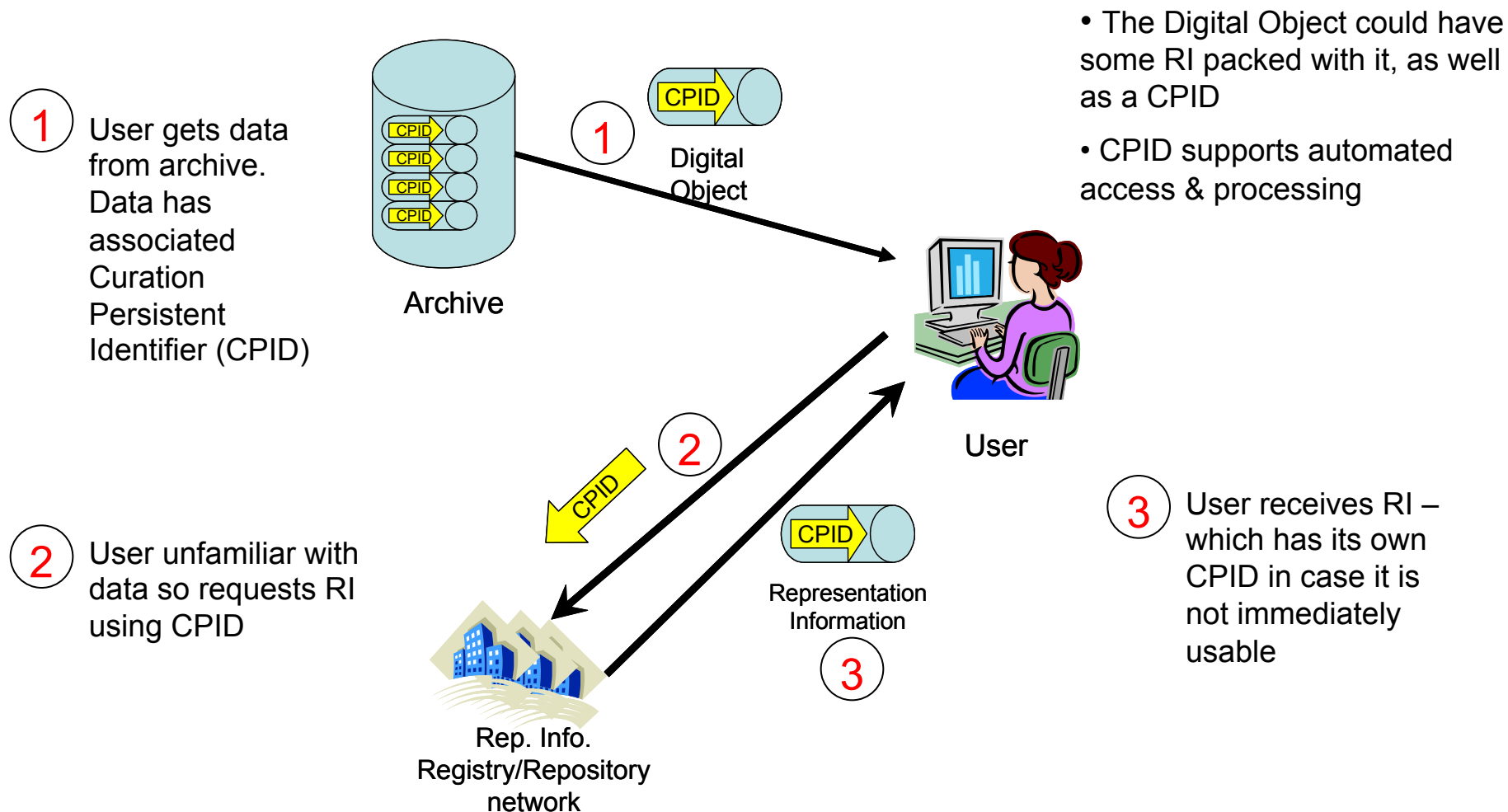
Registry/Repository of RI (RRoRI)

- Development started under the DCC-Development team
- Work now being undertaken jointly with the CASPAR Project
 - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (Integrated Project co-funded by EU FP6 Programme, April 2006)
- Representation Information is the key to long-term access
- RRoRI should itself be a trustworthy OAIS
- Repository: some RI is stored; Registry: links to external RI
- Emphasis on interoperability and automated use
- Vision is to have a global, distributed network of RI
- Provide an infrastructure of reliable and trusted RI for third party use

RRoRI: RI Label & CPID

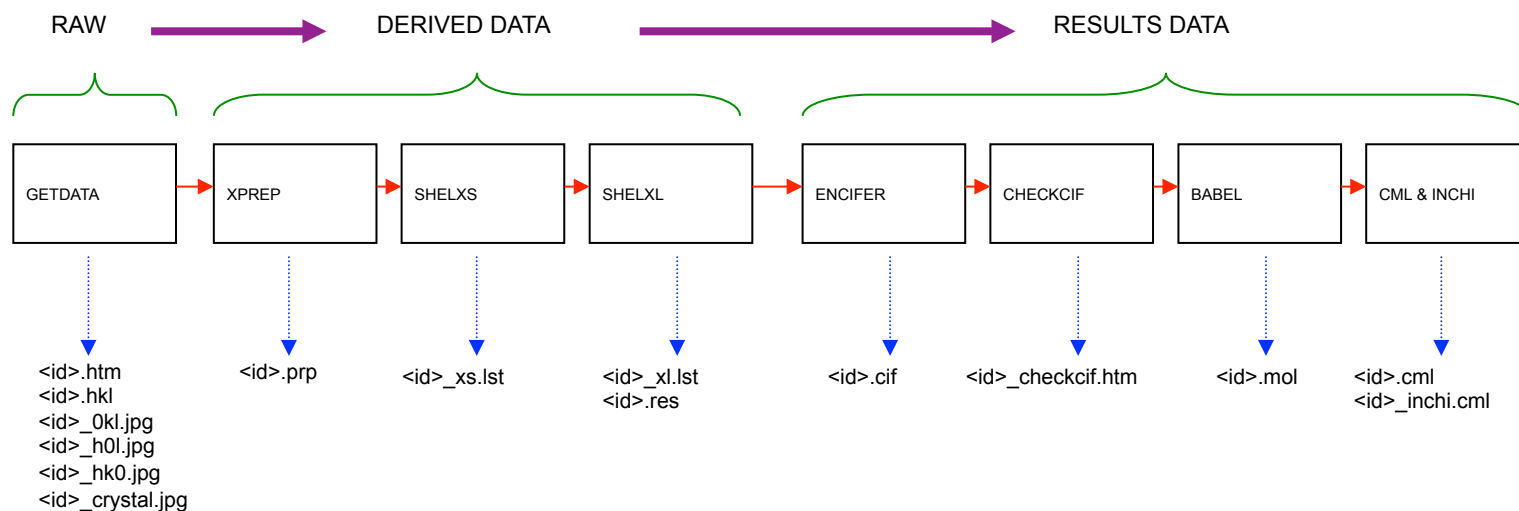
- Idea of RI is the key
 - **Information Object**: a specific object to be archived/preserved/curated
 - **RI**: all information required to render, interpret, use and understand the object
 - **RI Label**: used to connect RI to an Information Object
- RI Label serves as a mechanism for accessing RI in RRoRI
 - Label is used to identify relevant RI
 - Provides mechanism for recording individual RI components
- RI Label has a Curation Persistent Identifier (CPID)
 - Used to connect the digital object to the RI Label

Use of CPID

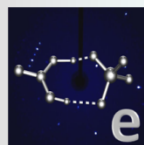
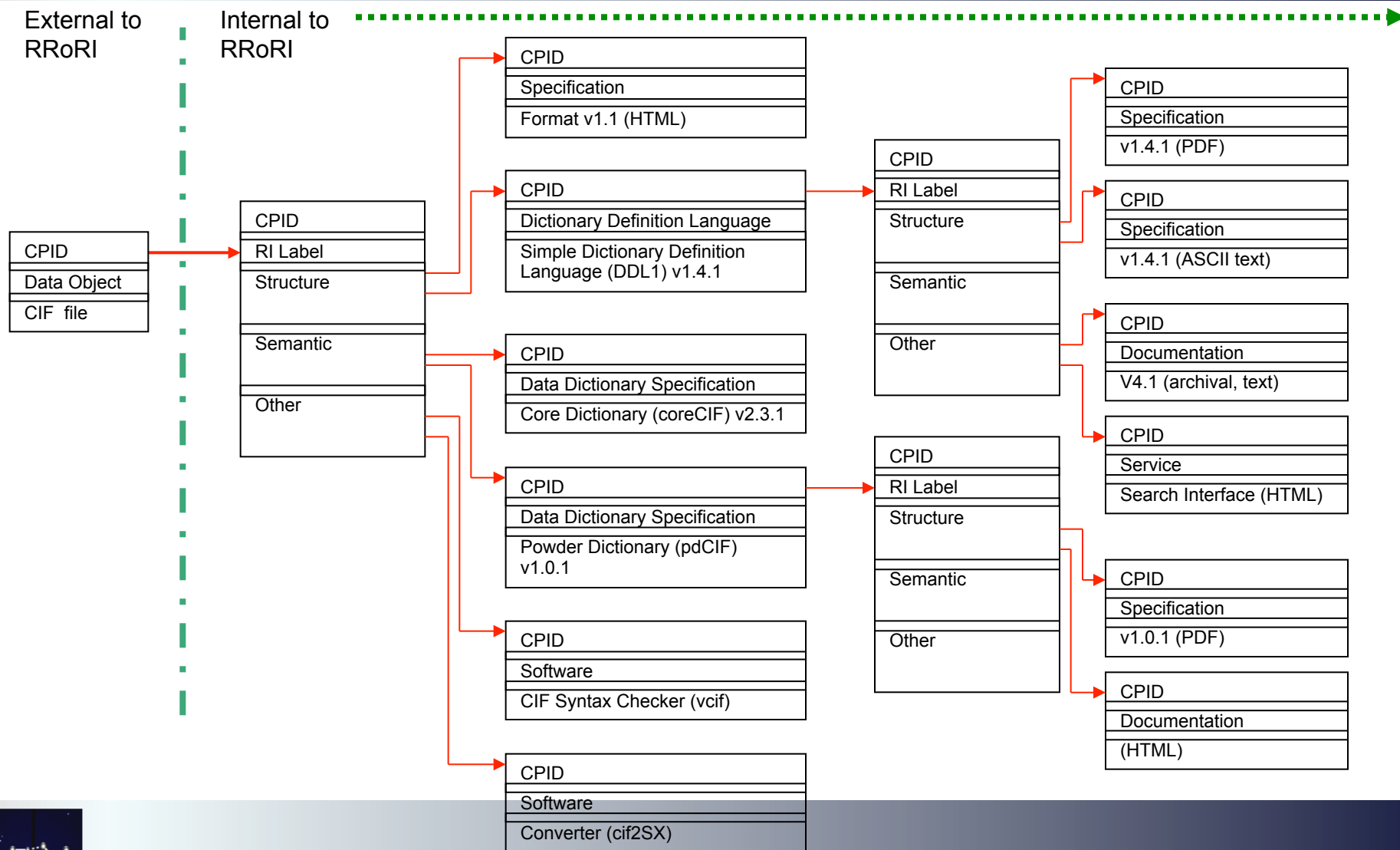


Capturing RI: Crystallography Data

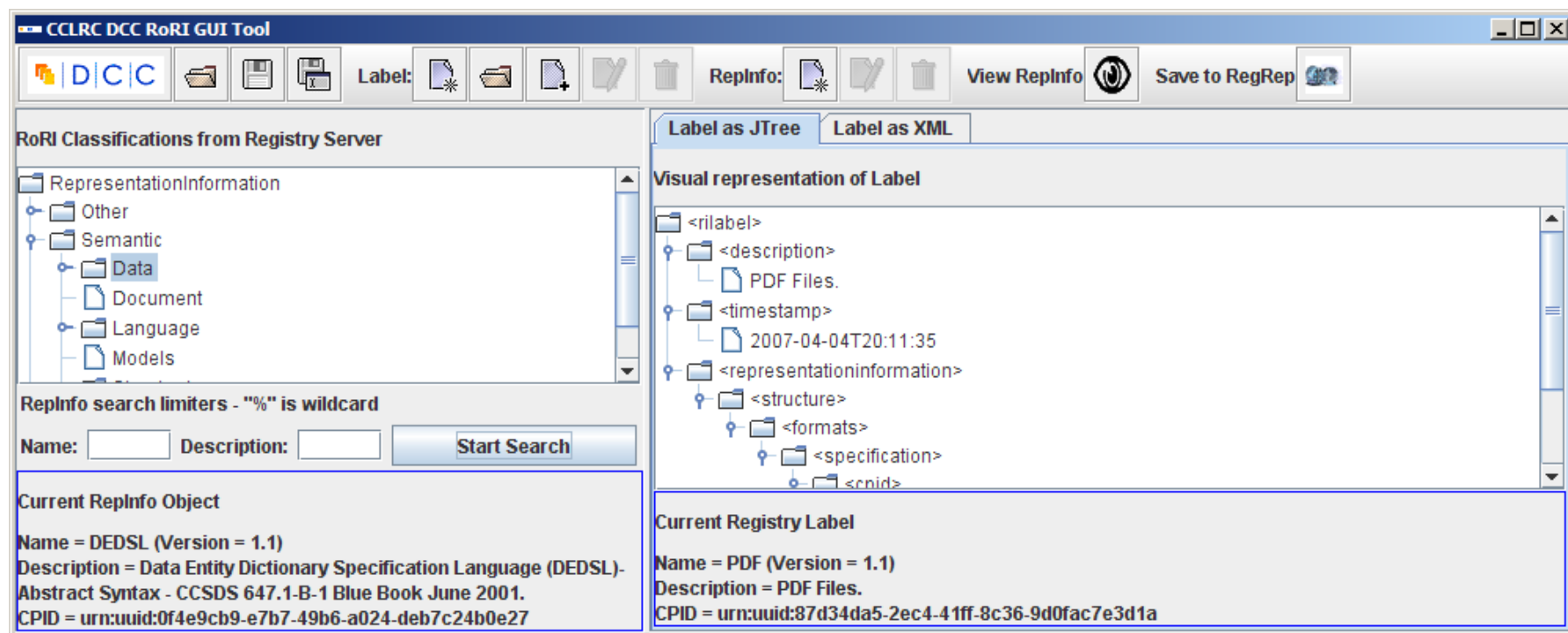
- Crystallography data are highly structured
- Convention is to share derived or reduced data, access to raw data is rare
- Crystallography Information File (CIF) is a de facto exchange standard
- CIF maintained by International Union of Crystallography (IUCr)
- Open standards and software e.g. CIF, checkcif, CML, INChI
- Culture for sharing/depositing data (CCDC)
- Well-established workflow for crystallography experiments



CIF RI Network (Partial)



RRoRI GUI Tool



A client with a GUI for ingest, search and retrieval of RI and RI labels

Conclusions

- A preservation strategy based on RI depends on a global, well-engineered, distributed infrastructure of RI
 - Needs coordination, collaboration and globally shared effort
 - Mining of RI networks for inference purposes
- Creation of robust RI networks requires domain expertise
- Likely to be gaps in global networks of RI
 - Business case for using a store of RI is clear, however the case for submitting RI to the global effort is less clear (commercial, IPR etc.)

Thank You

Questions?

Manjula Patel, Simon Coles

eCrystals Federation Project

http://wiki.ecrystals.chem.soton.ac.uk/index.php/Main_Page